NAZARBAYEV UNIVERSITY

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

ELCE 311: Interdisciplinary Design Project

Integrating Deep Learning and Vision Transformer Models for Pulmonary Nodule Segmentation and Classification

Final Report

Student name: Dosbol Erlan
Supervisor name: Prof. Muhammad Tahir Akhtar
The course Instructor: Sultangali Arzykulov

April 19, 2024

Contents

1 Abstract

Significant improvements in deep learning techniques in computer vision have proven to be powerful tools in early lung cancer detection. This paper presents end-to-end data pipeline and two part deep learning models for segmentation and classification of CT scan data to detect malignant pulmonary nodules. Both models are instances of Convolutional Neural Networks (CNNs) [2] and deep learning heuristics for regularization, weight initialization and preventing vanishing/exploding gradients were used. The models were trained on LUng Nodule Analysis (LUNA) dataset, comprising of 888 CT scans from 888 patients. We used U-Net architecture [13] for semantic segmentation of the CT scan data and used an architecture vaguely similar to VGG network [7] for classification model. Preliminary results indicate promising potential for automated diagnosis, with a strong foundation laid for future exploration. The paper charts a course for future work, including the examination of diverse segmentation models, deeper CNN architectures like ResNet [11] and DenseNet [12], and innovative approaches such as Vision Transformers (ViT) [5] to refine accuracy and revolutionize image analysis in medical diagnostics.

2 Introduction

In Kazakhstan, as is the case globally, lung cancer represents the most prevalent form of cancer [1], dominating the oncological landscape both in incidence and mortality. One of the reasons why lung cancer is so lethal is that detecting cancer is challenging due to its rarity in general population and severe consequences in case of late diagnosis. The fact that it is so infrequent among the population might lead to oversight by even experts in the field. Therefore, introduction of Deep Learning models for early detection of lung cancer might be a crucial innovation for oncology. Moreover, there is an urgent need for enhancements in cancer detection capabilities around the world because of critical shortage of specialists. Deep Learning models, made more accessible by frameworks like PyTorch, have been making progress in Computer Vision and especially in Object Detection since the AlexNet won the ImageNet Challenge in 2012 [2]. This breakthrough highlighted the capabilities of Convolutional Neural Networks (CNNs) [3]. CNNs have orders of magnitude less number of parameters than Multilayer Perceptrons [3]. Furthermore, CNNs have the property of translation invariance [3] as its inductive bias, which is one of the reasons why this particular structure was the most efficient on the spatial data (images, videos etc.) until recently. For the last decade, this shift in paradigm, advancements in parallel computational power and surge in Deep Learning research have made it possible to train deeper models on more complex data.

In 2017, Google researchers published paper on new Deep Learning architecture called "Transformer" [4] that was originally developed for machine translation. But further exploration of the architecture has shown that the Transformer architecture can be efficiently trained on many different types of data, including spatial data [5]. Following invention of Attention mechanism [4] and Vision Transformers (ViT) [5], the deep learning and computer vision community have seen significant shift towards the new architectures as they performed better in image classification competitions like ImageNet [6].

In this project, the Deep Learning models train using LUng Nodule Analysis (LUNA) dataset. LUNA dataset is derived from the larger LIDC-IDRI (Lung Image Database Consortium and Image Database Resource Initiative) database. Total size of LUNA dataset in compressed form is about 60GB and it consists of high resolution CT scans. The LUNA dataset has about 888 scans from 888 patients. Each CT scan was annotated by four different experts. Annotations include the location of the nodule, it's diameter, whether it is benign or malignant among many other features.

3 Background

3.1 Literature Review

Since AlexNet's [1] inception, CNNs have evolved significantly. The VGG (Visual Geometry Group) model by Simonyan and Zisserman introduced the concept of "blocks" of convolutional layers, allowing for deeper architectures [7]. The Network in Network (NiN) architecture furthered this by embedding micro neural networks within convolutional layers, reducing parameter count and enhancing model abstraction [8]. GoogLeNet introduced multiple branches within its architecture, effectively handling various scales of processing and improving gradient flow [9].

The development of Batch Normalization by Ioffe and Szegedy marked another advancement, normalizing layer inputs to accelerate training and mitigate vanishing gradients [10]. Residual Networks (ResNets) introduced residual connections, enabling the training of even deeper networks by improving feature reuse [11]. Finally, DenseNets built upon ResNets, further enhancing feature propagation and reuse, which led to more efficient networks with fewer parameters [12].

Different segmentation models were studied. U-Net, introduced by Ronneberger et al. [13], revolutionized biomedical image segmentation with a symmetric architecture enhancing localization and context usage, designed for efficient training with minimal data. V-Net by Milletari et al. [14] extended these principles into 3D, adding residual connections for better gradient flow in volumetric image segmentation. Çiçek et al.'s 3D U-Net adapted the U-Net for 3D datasets [15], focusing on learning from sparsely annotated data, making it suitable for detailed medical imaging tasks.

Vaswani et al.[4] revolutionized sequence processing by introducing the Transformer architecture, which relies solely on self-attention mechanisms. Dosovitskiy et al. [5] adapted Transformers for image recognition by treating image patches like words in NLP. Touvron et al. [16] enhanced data efficiency in image Transformers using distillation and attention recalibration. Liu et al. [17] proposed the Swin Transformer, employing shifted windows for hierarchical representation, while He et al. [18] introduced masked autoencoders for scalable self-supervised learning in vision tasks. Finally, Li et al. [19] refined the Multiscale Vision Transformer to improve classification and detection capabilities.

As per textbooks, two books were reviewed [20],[21] and they form the foundation for project's theoretical and practical knowledge in Deep Learning and Computer Vision.

3.2 Key Related Works

One of the key works that affected the current model is the VGG architecture [7] as a structure similar to it was implemented for the classification model. Similar to VGG architecture, the project model utilizes convolutional blocks instead of layers. Another paper model that was utilized in this project is the U-Net segmentation model [13]. This model was used in the project because it was the easiest to implement out of three segmentation models studied [13], [14], [15].

3.3 Beneficiaries

- **Patients and Doctors**.
- **Deep Learning and Computer Vision Technical Community**. This project contributes to the ongoing research and development within the deep learning and computer vision communities by addressing real-world problems.

4 Current Model

The hight level structure of the project is as follows:

1. **Data Loading.** Preprocessing and making mini batches of CT scan data for further processing.
2. **Segmentation.** Use of off-the-shelf U-Net segmentation model to do semantic segmentation of the CT scan to mark nodules.
3. **Classification.** Classify segmented modules into benign or malignant.
4. **Diagnosis.** Expert uses the output of the model for diagnosis.

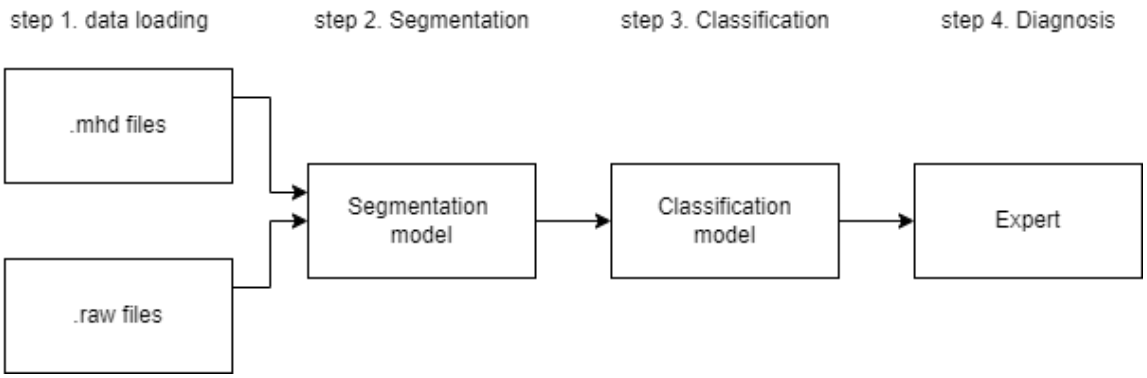


Figure 1: End-to-end structure of the Project

4.1 Data Loading

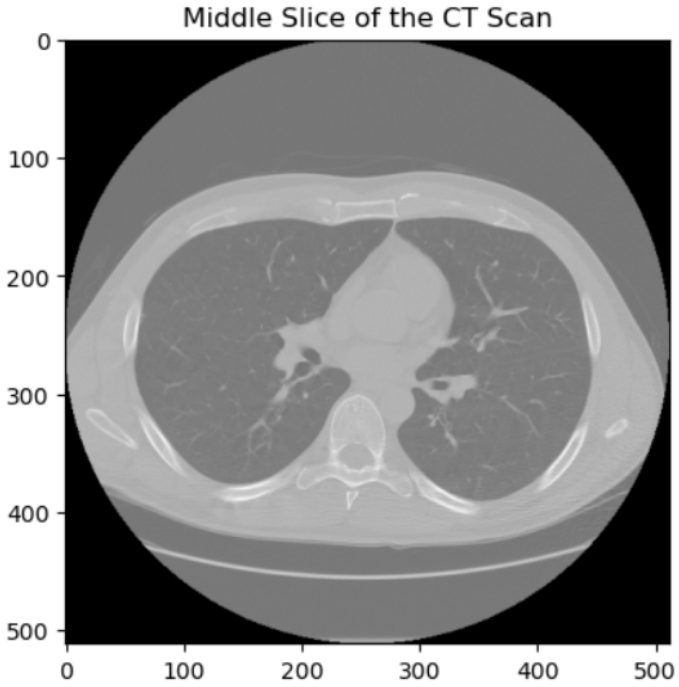


Figure 2: Sample slice of CT scan from LUNA dataset

CT images are three dimensional, therefore, we can visualize a slice of it as in Figure 2. CT images do not operate on conventional XYZ dimensions, but IRC (Index, Row, Column) dimensions. IRC dimensions are slightly different than XYZ dimensions because aspect ratio between IRC dimensions are not exactly 1:1:1. Therefore, we can see slight warping effect on the Sample Slice image from CT scan in Figure 2.

In LUNA dataset, there are three types of files:

- **.mhd files:** Metadata for the CT scan images such as dimensions, data type, modality etc.
- **.raw files:** Actual image data in unprocessed binary format.
- **.csv files:** Annotations and labels from experts.

Metadata files and raw image files are combined for CT array (Figure 3) and then Dimension Transformer is applied to change the dimensions of the CT array from IRC to conventional XYZ. The reason for this transformation is the fact that annotations are done in conventional XYZ dimensions. After the dimensions of the annotations and CT array no longer conflicting, we can combine them to create the Annotated Dataset. We utilize PyTorch built-in Data Loader to create mini batches of data for the segmentation model.

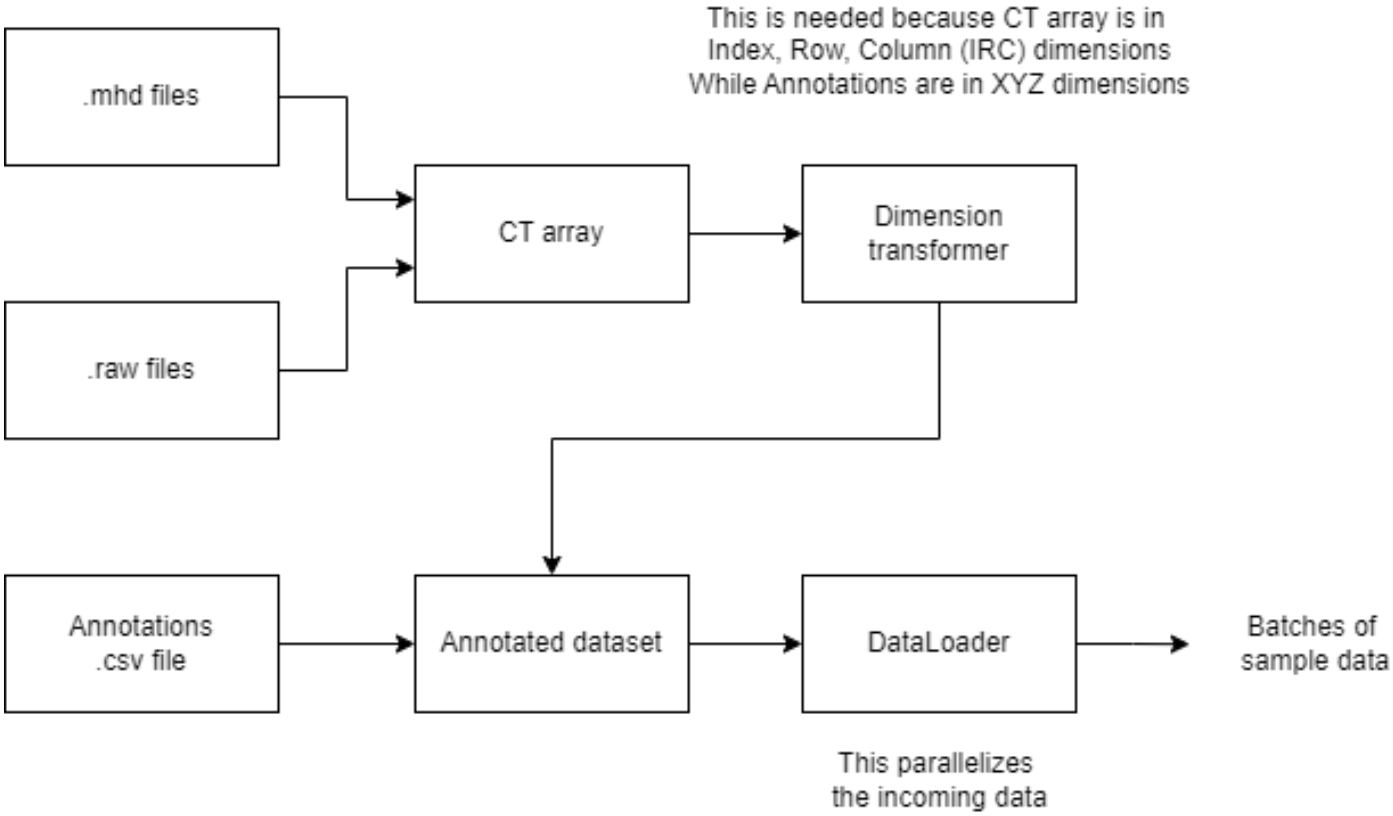


Figure 3: Data Loader Pipeline

4.2 Segmentation

For the segmentation model, we use U-Net [13] Model and consider it as black box as U-Net model is one of the ready-to-use models from the *torchvision* library. Input of the segmentation model is mini batches of CT images. Output of the segmentation model is mini batches of CT images with a binary mask applied where the sample space for the mask is {benign, malignant}.

4.3 Classification

The Classification architecture uses similar approach to designing the network as the VGG architecture [7] in a sense that it uses Convolution Blocks, not layers. Input of the classification model is the nodules marked by segmentation model. Output of the classification model is a set of pseudo probabilities of given nodules being benign or malignant. The structure of the model is as follows:

1. **Stem.** 3D Batch Normalization function [10] is applied to the input data to preprocess it, normalize it to prevent Vanishing or Exploding Gradients and regularize it to fight overfitting.

2. **Body.** Four CNN blocks (Figure 5) are connected in series. The idea is to increase the number of channels, while decreasing the volumetric resolution of the nodule image. One channels enters the CNN Block 1 and 64 channels are the output of the last CNN block. On the other hand, the volumetric resolution of the nodule image decreased from 32x48x48 to 2x3x3. Convolutional layers increase the number of channels, while MaxPooling functions decreate the volumetric resolution. The first CNN blocks learn low level feature (e.g. edges), while last CNN blocks capture high level features (e.g. shape) of the nodule image.

3. **Head.** The output of the body is not 1 dimensional, but 4 dimensional if we ignore the mini batch dimension. Therefore, we "flatten" the nodule feature maps into 1 dimensional vector. The vector is sequentially goes through a Fully Connected Layer and SoftMax function, which finally outputs pseudo probability for whether given nodule is bening or malignant.

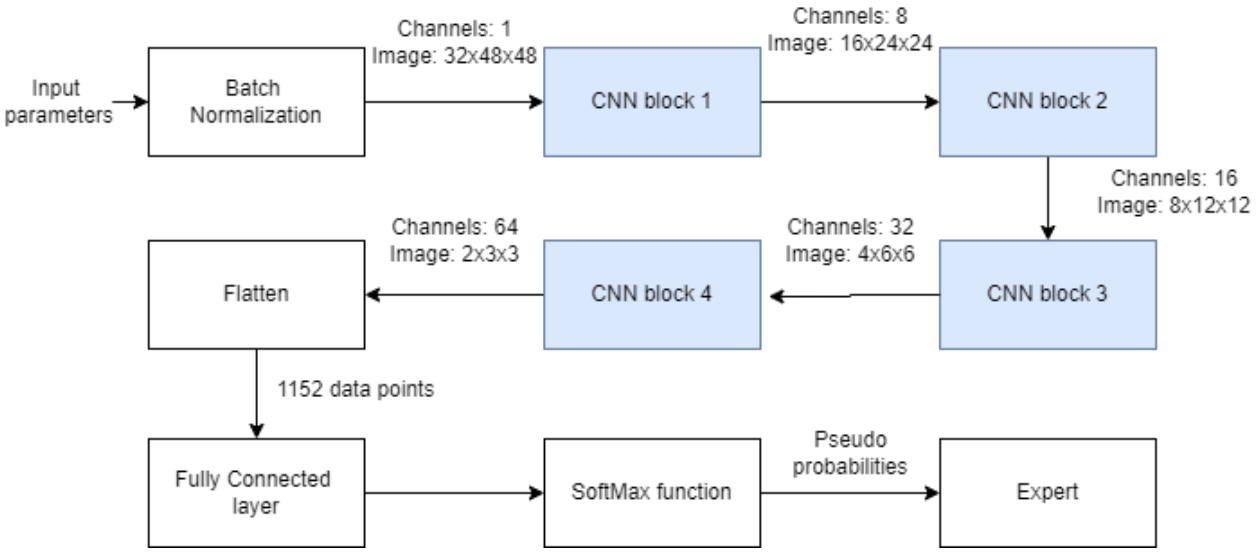


Figure 4: Classification model architecture

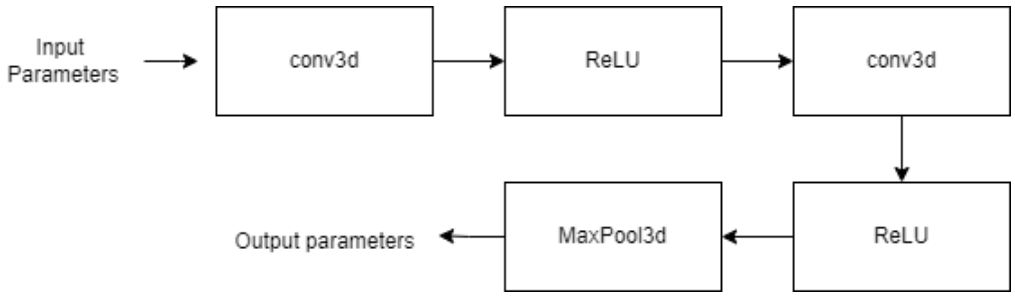


Figure 5: CNN block

4.4 Diagnosis

The Diagnosis stage is carried out by expert based on their experience, patient's health record among many other factors. The cancer detection model cannot replace an expert, but it can be regarded as one of their tools.

5 Preliminary Results

In Table 1, the validation results of the classification model is presented. Total of 161 nodules were in the validation set and 12 of them were missed by the Segmentation Model (92 % success rate). The classification model filtered out 9 % of the actual nodules and 98 % of the non-nodules. Further results, are given below in Table 2, Table 3 and hyperparameters values for two models are given in Table 4 and Table 5.

Table 1: Summary of Detection and Classification Results

Total	Missed by Segmentation	Filtered Out by Classification	Pred. Benign	Pred. Malignant
Non-Nodules		167015	1482	452
Benign	11	4	73	19
Malignant	1	8	10	35

Table 2: Confusion Matrix - Interest in Malignancy

	Predicted Malignant	Predicted Benign
Actual Malignant	35 (TP)	10 (FN)
Actual Benign	19 (FP)	73 (TN)

Definitions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 3: Performance Metrics

Metric	Value
Accuracy	0.7883
Precision	0.6481
Recall	0.7778
F-Score	0.7071

Parameter	Value
Training epochs for nodule/non-nodule classification	100
Training epochs for benign/malignant classification	40
Fine-tuning depth	2 layers
Optimizer	Stochastic Gradient Descent
Learning Rate	0.001

Table 4: Training hyperparameters for classification models.

Parameter	Value
Training epochs for segmentation model	15
Optimizer	Adam
Learning Rate	0.001

Table 5: Training hyperparameters for segmentation model.

6 Project Implementation Plan

6.1 Project Plan for Capstone I

- **Theoretical Study of CNNs:** Study different CNN architectures and implement them from scratch in PyTorch deep learning framework.
- **Literature Review:** Review and experiment with VGG[11], NiN [12], GoogLeNet[3], ResNet[2], DenseNet[14] and other architectures.
- **Model Implementation and Training:** Use one of the reviewed architectures or come up with some custom architecture for classification and/or segmentation of the CT data from LUNA dataset.
- **Testing and Evaluation of Results:** Test the results of the trained model with benchmark results trained using different techniques.

6.2 Project Plan for Capstone II

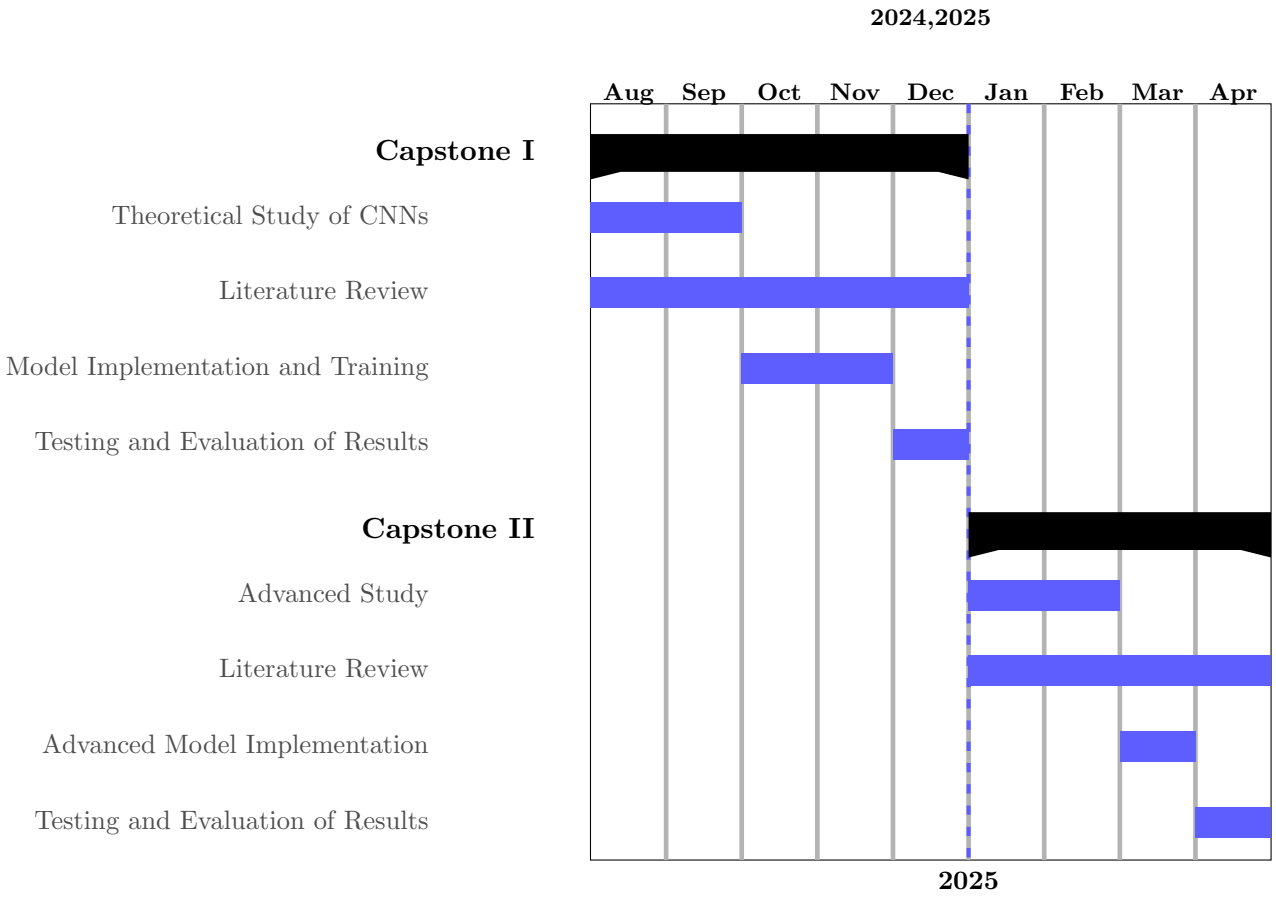
- **Advanced Study:** Study different Transformer architectures and implement them from scratch in PyTorch deep learning framework.
- **Literature Review:** Review seminal ViT papers [16], [17], [18], [19], [20] and experiment with them.
- **Advanced Model Implementation:** Use one of the reviewed architectures or come up with some custom architecture for classification and/or segmentation of the CT data from LUNA dataset.
- **Testing and Evaluation of Results:** Test the results of the trained model with benchmark results trained using different techniques.

6.3 Risk and Contingency Plan

Risk	**Contingency Plan**
There might be bottlenecks in the data pipeline, which would lead to not meeting the deadline	A lot of effort will be put into optimization of the data flow and utilize cloud parallel computing services.
Memory and compute requirements for the model might exceed the budget	Smaller datasets and architectures will be used instead

Table 6: Risk and Contingency Plan Table

6.4 Gantt Chart



6.5 SWOT Analysis

- **Strengths**: Availability of Open Source Learning Materials, Deep Learning frameworks, Datasets and Models.
- **Weaknesses**: Medical Datasets tend to be huge in size, but few in examples and difficult to obtain due to Patient Privacy.
- **Opportunities**: To develop new architecture or heuristics for the segmentation, classification, pre-processing of the CT data.
- **Threats**: To misdiagnose a person with/without malignant nodules. Both outcomes are undesirable.

7 Conclusions and Future Work

Conclusions

Our evaluation of the model's performance on lung nodule detection and classification yielded notable outcomes:

- The model was successfully trained on an extensive dataset, totaling 60GB, demonstrating the scalability of our approach.
- Integration of the U-Net [13] architecture facilitated effective segmentation, underscoring the utility of proven methodologies in medical image analysis.
- Preliminary results are positive, indicating the model's potential in accurately identifying lung nodules, a critical step towards automated diagnosis.

Future Directions

In pursuit of refining our model's accuracy and reliability, we outline several avenues for future research:

- **Segmentation Model Exploration:** Research and experiment with diverse segmentation models and architectures [14] , [15], further broadening the scope and applicability of our work in the domain of medical imaging and beyond.
- **CNN Architectures:** Experiments with robust architectures like ResNet [11] and DenseNet [12] are planned, as their ability to facilitate deeper neural networks could prove invaluable.
- **ViT Architectures:** Novel approaches like Vision Transformers [5], [16], [17], [18], [19], which depart from traditional convolution-based methods, will be investigated for their potential to revolutionize image analysis tasks.

References

[1] D. Yessenbayev et al., "Epidemiology of Lung Cancer in Kazakhstan: Trends and Geographic distribution," *Asian Pacific Journal of Cancer Prevention*, vol. 24, no. 5, pp. 1521–1532, May 2023, doi: 10.31557/apjcp.2023.24.5.1521.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097-1105, 2012.

[3] Y. LeCun and Y. Bengio, Convolutional networks for images, speech, and time series. 1995, pp. 255–258. *The handbook of brain theory and neural networks*

[4] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 6000-6010, 2017.

[5] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. on Learning Representations*, 2020.

[6] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, doi: 10.1109/cvpr52688.2022.01179.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[9] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1-9, 2015.

[10] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016.

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, pp. 234-241, Springer, 2015.

[14] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth Int. Conf. on 3D Vision (3DV)*, pp. 565-571, IEEE, 2016.

[15] Ö. Çiçek et al., "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, vol. 9901, pp. 424-432, Springer, 2016.

[16] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in *Proc. of the 38th International Conference on Machine Learning*, PMLR, 2021.

[17] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2021.

[18] K. He et al., "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[19] Z. Li et al., "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.